Personalizing and Improving Tag-Based Search in Folksonomies

Samia Beldjoudi, Hassina Seridi-Bouchelaghem, Catherine Faron Zucker

Personalizing and Improving Tag-Based Search in Folksonomies

Samia Beldjoudi[1], Hassina Seridi-Bouchelaghem[1], and Catherine Faron-Zucker[2]

[1] Laboratory of Electronic Document Management LabGED,
Badji Mokhtar University Annaba, Algeria
{beldjoudi,seridi}@labged.net
[2] I3S, Université Nice - Sophia Antipolis, CNRS 930 route des Colles, BP 145,
06930 Sophia Antipolis Cedex, France
catherine.faron-zucker@unice.fr

Abstract. Recently, the approaches that combine semantic web ontologies and web 2.0 technologies have constituted a significant research field. We present in this paper an original approach concerning a technology that has recognized a great popularity in these recent years, we talk about folksonomies. Our aim in this contribution is propose new technique for the Social Semantic Web technologies in order to see how we can overcome the problem of tags' ambiguity automatically in folksonomies even when we choose representing these latter with ontologies. We'll also illustrate how we can enrich any folksonomy by a set of pertinent data to improve and facilitate the resources' retrieval in these systems; all this with tackling another problem, we speak about spelling variations.

Keywords: Folksonomies, Web 2.0, Semantic Web, Tags Ambiguity, Spelling Variations.

1 Introduction

Among the powerful technologies of Web 2.0, we find folksonomies, this term has recently appeared on the net to describe a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content. Ontologies which constitute the backbone of semantic web contribute significantly in solving the problems of semantics during the definition and the search of information. However even with the strong points of folksonomies and ontologies; their combination together still suffers from some problems. As examples we can cite the problem of tags' ambiguity and spelling variations (or Synonymy) in folksonomies. Our goal in this contribution is to show how we can exploit the power of social interactions between the folksonomy's members in order to extract the meaning of terms and overcome the problems of tags' ambiguity and spelling variations. Also we will try to show how we can use the principle of rules-based systems with ontologies for helping our system to enhance automatically the folksonomy by relevant facts can increase the data available within our system with

relevant information for facilitating the resources retrieval and optimizing the time expended during this process. Our paper is organized as follows: Section 2 presents a quick overview about the main contributions attached to our search field; in Section 3 we will detail the design of our approach. After in Section 4 we move to the experimental phase in order to measure the performance of our approach and discuss the obtained results. Conclusion and future works are discussed in Section 5.

2 Related Work

In this section, we will put the point on the famous works which try to reduce the tags' ambiguity problem and especially those aimed to extract the semantic links between folksonomy's terms using ontologies. Mika [7] proposed to extend the traditional bipartite model of ontologies to a tripartite one: that of folksonomies. In another work, Gruber [5] argued that there is no contrast between ontologies and folksonomies, and therefore recommended to build an "ontology of folksonomy". According to Gruber, the problem of the lack of semantic links between terms in folksonomies can be easily resolved by representing folksonomies by ontologies. Specia and Motta [9] in their turn have preferred the use of ontologies to extract the semantics of tags. Their approach consists in building tags clusters, and then trying to identify the possible relationships between tags in each cluster. The niceTag project of Limpens et al. [6] is focused on this same principle: the use of ontologies to extract semantic links existing between tags in a system. In addition, this project has introduced the idea of exploiting interactions between users and the system. Pan et al. [8] aimed at reducing the problem of ambiguity in tagging. They proposed to extend the search of tags in a folksonomy by using ontologies. They defended this principle of extension of the search in order to avoid bothering the users with the rigidity of ontologies. Beldjoudi et al. [1] proposed a technique specially designed to show the social interactions' usefulness in folksonomies for reducing tags' ambiguity problem. In another contribution the one of Beldjoudi el al. [2], the authors propose a method to analyze user profiles according to their tags in order to personalize the recommendation of resources. To sum up, most of the works relative to folksonomies aim to bring together ontologies and folksonomies as a solution to the tags' ambiguity problem and that of the lack of semantic links between tags. In this context, we started our trial to improve a little this technology and give a new view concerning the combination between folksonomies and ontologies.

3 Semantic Social Folksonomy with Ontology (SSFO)

Our aim in this contribution is to introduce both the semantics and the social aspects in folksonomies in order to let any user in the system retrieving relevant web resources close to his preferences. In this paper, we aim to show how we can produce a technique for helping any ontology already used for representing a folksonomy to overcome the problem of tags' ambiguity automatically without the need of an expert who must control and organize links between terms. In addition we want show how

we can enrich our folksonomy (without human intervention) with relevant data in order to help optimizing the time of search and enormously reduce the problems of spelling variations and the lack of semantics within folksonomies focusing on the rules-based systems.

3.1 Formal Description

Formally, a folksonomy is a tuple **F = <U, T, R, A>** where **U**, **T** and **R** represent respectively the set of users, tags and resources, and **A** represents the relationship between the three preceding elements i.e. **A ⊆ U x T x R**. Because this approach is intended to present a technique that can help any folksonomy represented by an ontology to overcome the problems of tags' ambiguity and spelling variations based on the preferences and the interests of each user, and also enrich automatically the system by new relevant data, we suggest here to represent our folksonomy with a simple ontology defined by primitives relations such as "tagged by" and "used by"… etc.

3.2 Resolving Tags' Ambiguity in Folksonomies

Our technique to overcome the problem of tags' ambiguity is not based on ontologies. The idea is to study the profile of each member in the system and then compare the preferences of this one with other users in order to extract those who are similar to him.

It should be noted that: To make the system flexible, we propose to make it interact with the user to accept or reject the retrieved resources. And to avoid the "cold start" problem which is generally occur from a lack of the required data by the system in order to make an excellent recommendation; it's proposed to measure the similarity between resources when the users are not similar. So we can summarize our methodology as follow:

Similarities between Users. To calculate this similarity we suggest to use a measure that allows representing each user by a vector v_i designates a series of binary numbers defined the set of his tags or his resources. Thus, to calculate the similarity between two users, for example U_1 and U_2, this measure proposes to calculate the cosines of the angle between their associated vectors v_1 and v_2 as shown in the formula (1):

$$cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|^2\|v_2\|^2} \quad (1)$$

Similarities between Resources. When the users are not similar we suggest measuring the degree of similarity between resources in order to avoid "cold start" problem which is generally resulted from a lack of the data required by the system in order to make an excellent recommendation.

Recommendation Levels. We propose here assigning to each resource recommended by the system a factor that indicates the percentage of its recommendation. To achieve this classification, we propose to calculate the ratio between the number of resources

used by the user himself (i.e. the one who does the search) and the number of the resources shared between him and the other users. Above a threshold fixed in [0..1], we qualify the resource as *highly recommended*; under this threshold, it is simply *recommended* or *weakly recommended* if the similarity is close to zero.

3.3 Rules-Based Systems in Folksonomies

The purpose of using rules-based systems can be summarized as follow: 1) Avoid the existence of an expert who must control and organize links between terms. This let us say that our technique is dynamic and automatic. 2) Optimize and reduce the time required for searching relevant resources for each user by avoiding the recalculation of similarities every time. And 3) enrich the folksonomy by a relevant fact which can help improving the process of search and reducing the problem of spelling variations. In our approach the folksonomies' enrichment is realized by two categories of data as follows:

1. Enrich our fact base by facts extracted from the similarities' calculations that have been made during the step 3.2; and which say that: such resource is similar to such resource. For example; if we have already found that a resource R1 is similar to another resource R2, then we can add the following fact: is-similar-to (R1, R2) which express that *"R1 is similar to R2"*. With this method our system does not recalculate the similarity between the users every time when an actor want to search relevant resources, but it will optimize this time and also the memory space that can be lost in each calculation because with this process; before our system begin the calculation of similarity between users or between resources it will firstly see in the fact base if there are resources similar to those already proposed to this user.
2. The second kind of facts has the following form: "A resource R_Z can have as tags the tag T_Y" or can-tagged-by (R_Z, T_Y). The advantage of such fact is twofold: a) Reduce the problem of tags' ambiguity (because the similarity between resources became more highly). b) Reduce the problem of spelling variations. We can explain this second point (b) by the following example: "cat" and "chat" means both the same concept (animal) in English and in French, but when a user searches resources annotated by the tag "cat", the system will not offer him those tagged by the word "chat" because it can't understand that the tag "cat" is equivalent to the tag "chat". In others words, supposing that the user U_X tagged a resource R_1 by the tag *cat* and U_W is the user who tagged the resource R_2 by the tag *chat*. Noting that; the two resources R_1 and R_2 are already considered as similar according to the similarities' calculations that have been made before. Now if the user U_X wants search resources concerning the animal "cat" by the tag *cat*, the resource R_2 will not be given to him. In order to overcome this problem our approach proposes to add the following facts: can-tagged-by (R_1, chat), can-tagged-by (R_2, cat). And now any user can benefit from the resources of the other and so we have overcame the problem of spelling variation in folksonomies.

Finally, the relationship between our solution and the above problems can be appeared behind the choice of the rules' language RIF (Rule Interchange Format), which became recently a W3C Recommendation. The choice of this language is motivated by the fact that it can support the import of RDF data and RDFS/OWL ontologies [3]. Also a mapping to RIF from ontologies and the vice versa is possible, and thus we can easily treat our dataset and enrich the folksonomy. Furthermore the strength of this language is appear from the fact that it can support many kinds of dialects; among them we find the RIF-PRD (the Production Rule Dialect of the W3C Rule Interchange Format) [4] which allows adding, deleting and modifying facts in the fact base. And so can modify, assert and also retract a set of facts in our data base according to our needs.

4 Experimentation

4.1 Dataset and Data Treatment

In order to validate our approach, we have conducted an experiment with del.icio.us database. Our test base comprises 1605 tag assignments involving 55 users, 526 tags some of which are ambiguous or have many spelling variations, 950 resources each having possibly several tags and several users. To demonstrate the validity of our approach, we have distinguished two classes of users: the first one contains the users who have employed ambiguous tags and the other one those who did not use those tags. This ambiguity of tags has been subjectively decided: for instance *apple* is ambiguous and *software* is not.

First of all, we have constituted a simple ontology from this dataset in order to represent the folksonomy by ontology. It should be noted that we have used a simple properties for describing this ontology in order to avoid losing the meaning and the objective of our approach, where we have suggested representing our folksonomy by a simple ontology defined by primitives relations such as "tagged by" and "used by"…etc. After that we have used a tool for social network analysis called "Pajek[1]", in order to extract the three networks 'Users-Tags', 'Users-Resources' and 'Tags-Resources'. The results of this step are used in our methodology to calculate the similarities between users and between resources in order to detect the pertinent resources for each user. Now, once we have extracted the three social networks and calculate the different similarities, we have choosen to represent these data with the language RIF because it allow us representing and manipulating our data easily since it can manage RDF data and RDFS /OWL ontologies.

4.2 Results

Three metrics are used for evaluating our approach: Precision: It measures the system's ability to reject all not relevant resources to a query. It is given by the ratio of all relevant selected resources and the set of all selected resources. Recall: It measures

[1] It's an analytical tool of social networks, used in [7].

the ability of the system to retrieve all relevant resources to a query. It is given by the ratio of relevant retrieved resources and all relevant resources in the database. And the metric F1: Which is a combination of the two previous metrics and is defined by the formula (2): $F_1 = \frac{2 . Precision . Recall}{Precision+Recall}$ (2). The three metrics listed above are calculated for each user, and then the average of each metric in the system is calculated. The results are shown in Figures 1 and 2.

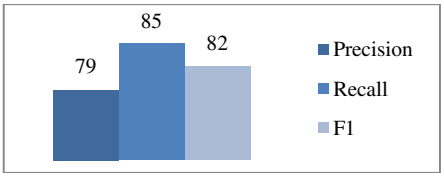

Fig. 1. The average of the three metrics concerning the problem of tags' ambiguity

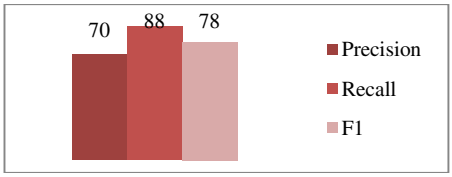

Fig. 2. The average of the three metrics concerning the problem of spelling variations

4.3 Discussion

The approach presented in this work has tried to extract the semantics in folksonomies in order to allow users capturing the social dimension of their tagging activity. Indeed the obtained results show that the technique SSFO succeeded in distinguishing between ambiguous tags and also them which have many spelling variations. Comparing now our approach with other ones trying to discuss the problem of tags' ambiguity; for example the Pan's and al work [8], we can conclude that our results are very optimistic especially when we know that the proposed approach is flexible i.e. the result of the search's procedure can be changed according to the profile and the interests of each user in contrary to the other approach. In addition the work presented in [8] doesn't tackle the problem of spelling variations. In comparison with [5], we find that our approach doesn't need an expert who must control and organize links between terms. Also the expertise of users which was introduced in [6] is characterized by the complexity of its exploitation when we try as much as possible to avoid a cognitive overload, to limit the necessary effort for the formalization of this expertise which is achieved by our approach.

5 Conclusion and Future Work

We have proposed in this paper a new technique based on the force of social interactions between the different actors in folksonomies in order to create a consensus among the users and so increase the semantics in these systems. We have tested this approach on a small amount of data and we have obtained good results. In order to expand and improve this work we propose to validate our approach on a larger dataset and also enrich our database by other relevant facts.

References

1. Beldjoudi, S., Seridi, H., Faron-Zucker, C.: Ambiguity in Tagging and the Community Effect in Researching Relevant Resources in Folksonomies. In: Proc. of ESWC Workshop User Profile Data on the Social Semantic Web (2011)
2. Beldjoudi, S., Seridi, H., Faron-Zucker, C.: Improving Tag-based Resource Recommendation with Association Rules on Folksonomies. In: Proc. of ISWC Workshop on Semantic Personalized Information Management: Retrieval and Recommendation (2011)
3. de Bruijn, J. (ed.): RIF RDF and OWL Compatibility. W3C Recommendation (June 22, 2010), http://www.w3.org/TR/2010/REC-rif-rdf-owl-20100622/
4. de Sainte Marie, C., Hallmark, G., Paschke, A. (eds.): RIF Production Rule Dialect. W3C Recommendation (June 22, 2010), http://www.w3.org/TR/2010/REC-rif-prd-20100622/
5. Gruber, T.: Tag Ontology-a way to agree on the semantics of tagging data (2005)
6. Limpens, F., Gandon, F., Buffa, M.: Sémantique des folksonomies: structuration collaborative et assistée, IC (2009)
7. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
8. Pan, J., Taylor, S., Thomas, E.: Reducing Ambiguity in Tagging Systems with Folksonomy Search Expansion. In: Proc. of the 17th International World Wide Web Conference (2009)
9. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)